Mühendislik Bilimleri ve Tasarım Dergisi
9(1), 52 – 63, 2021
e-ISSN: 1308-6693

Journal of Engineering Sciences and Design
DOI: 10.21923/jesd.671836

Araştırma Makalesi Research Article

AIR QUALITY INDEX PREDICTION IN BESIKTAS DISTRICT BY ARTIFICIAL NEURAL NETWORKS AND K NEAREST NEIGHBORS

Burhan BARAN[*]

Directorate of Environment and Urbanization, Malatya, Turkey

Keywords

Air Quality Index,
Artificial Neural Networks,
K Nearest Neighbors,
Classification.

Abstract

In this study, Air Quality Index (AQI) in Besiktas was intended to be predicted by Artificial Neural Networks (ANN) and k Nearest Neighbors (kNN) algorithms. For this purpose, eight parameters have been selected which may affect the AQI. These parameters are PM_{10}, SO_2, CO, O_3, temperature, humidity, pressure and wind speed, respectively. 124 data for 2015, 2016, 2017 and 2018 January, which includes eight parameters, were determined as training data. The first 14-day data of January 2019 were determined as test data. Similarly, the first 14-day data of January, March and December of 2018 were used as test data. In addition, The first 14-day data for January 2019 were normalized and set as test data. The success of ANN and kNN were measured by comparing. Performance rate of ANN with raw data for January 2018 was 85.71%, for March 2018 was 71.43%, for December 2018 was 78.57%. Both with raw and with normalized data for January 2019 was 85.71% performance rate. Performance rate of kNN with raw data for January 2019 was 92.86%, for March 2018 was 28.57%, for December 2018 was 71.43%. Performance rate of kNN with normalized data for January 2019 was 92,86%.

BEŞİKTAŞ'TAKİ HAVA KALİTESİ İNDEKSİNİN YAPAY SİNİR AĞLARI VE K EN YAKIN KOMŞULUK ALGORİTMALARI İLE TAHMİNİ

Anahtar Kelimeler

Hava Kalitesi İndeksi,
Yapay Sinir Ağları,
K En Yakın Komşuluk,
Sınıflandırma.

Öz

Bu çalışmada, Beşiktaş'taki Hava Kalitesi İndeksinin (HKİ) Yapay Sinir Ağları (YSA) ve k En Yakın Komşuluk (kNN) algoritmaları ile tahmin edilmesi amaçlanmıştır. Bu amaçla HKİ'yi etkileyebilecek 8 adet parametre seçilmiştir. Bu parametreler sırasıyla PM_{10}, SO_2, CO, O_3, sıcaklık, nem, basınç ve rüzgar hızı'dır. Bu parametreleri içeren 2015, 2016, 2017 ve 2018 yıllarının Ocak aylarına ait 124 adet veri eğitim verisi olarak belirlenmiştir. 2019 yılı Ocak ayına ait ilk 14 günlük veriler ile 2018 yılının Ocak, Mart ve Aralık aylarının ilk 14 günlük verileri ise test verisi olarak kullanılmıştır. 2019 yılı Ocak ayının ilk 14 günlük verisi normalize edililerek, ayrıca test verisi olarak ta kullanılmıştır. YSA ve kNN'nin sonuçları karşılaştırılarak başarıları ölçülmüştür. YSA'nın Ocak 2018 ham verileri ile başarım oranı % 85.71, Mart 2018 için % 71.43, Aralık 2018 için % 78.57 olmuştur. Hem ham hem de normalize edilmiş Ocak 2019 verileri için ise başarım oranı % 85,71 olmuştur. kNN nin Ocak 2019 ham verileri ile başarım oranı % 92.86, Mart 2018 için % 28.57, Aralık 2018 için % 71.43 olmuştur. kNN'nin normalize edilmiş Ocak 2019 verileri ile başarım oranı ise % 92.86 olmuştur.

Alıntı / Cite

Baran, B., (2021). Air Quality Index Prediction In Besiktas District By Artificial Neural Networks And K Nearest Neighbors, Journal of Engineering Sciences and Design, 9(1), 52-63.

Yazar Kimliği / Author ID (ORCID Number)	Makale Süreci / Article Process	
B. Baran, 0000-0001-6394-412X	**Başvuru Tarihi / Submission Date**	08.01.2020
	Revizyon Tarihi / Revision Date	28.12.2020
	Kabul Tarihi / Accepted Date	10.01.2021
	Yayım Tarihi / Published Date	30.03.2021

[*] İlgili yazar / Corresponding author: burhanbaran@gmail.com

1. Introduction

Everyone needs the environment to survive. In parallel with the rapid increase in the world population, irregular urbanization and industrialization have brought many environmental problems. Air pollution is defined as the chemical structure of the atmospheric layer is damaged by the mixing of various chemicals such as gas, dust and aerosol, causing damage to living things and even death. In other words, it is the presence of substances in the atmosphere that harm human and other living things. Air pollution can also be defined as a mixture of gases and solid particles in the air. Exhaust gas from vehicles, particles from solid fuels used in houses, chemicals from factories, dust and pollen are the main factors causing air pollution. In addition to investigating the causes of air pollution, it is also important to find solutions. Determining the sources and the levels of air pollutant has an important place in air quality control studies. Therefore, the prediction of pollution levels is an important issue in the regions where air pollution is present (Menteşe et al. 2012, Dursun et al. 2014, Kaplan et al. 2014, Kuntet al. 2018). Accordingly, many countries have established air quality management systems to monitor and control air pollution in cities (Alkasassbeh et al. 2013). Ambient air quality is primarily affected by particulate matter (PM), sulfur dioxide (SO_2), nitrogen dioxide (NO_2), ozone (O_3) and carbon monoxide (CO) concentrations in the atmosphere (Xieet al. 2015). PM_{10} pollutant is the representation of particulate matter. The particulate material is a mixture of solid particles and liquid droplets suspended in the atmosphere. It consists of a wide range of sources, including dust, smoke and soot, as well as visible and invisible particles. Fossil fuels, pollens, dust from the ground, dust from construction and industry can be give in as examples. The “10” in PM_{10} indicates particles with aerodynamic diameters less than 10 μm (Turgut et al. 2015). It can be found both outdoors and indoors. It can cause lung diseases, cardiovascular diseases, heart attack and rhythm disturbance (Karacı, 2018; Tepe et al. 2019). Sulfur dioxide is a gaseous pollutant. It is invisible and has a sharp smell. Facilities that burn fossil fuels are the main sources of sulfur dioxide. Sulfur dioxide is also present in motor vehicle emissions (Sulfur Dioxide, 2019). It is directly toxic to humans and affects respiratory functions. Carbon monoxide is a colorless, odorless gas that can be harmful when inhaled in large amounts. It can cause people to faint and die due to poisoning. The ozone molecule (O_3) is harmful to air quality, outside of the ozone layer (CO, 2019). It is known that there is a close relationship between air pollution concentrations and meteorological factors. Some of them are temperature, humidity, pressure and wind speed. These parameters do not have any negative effect on human health in terms of air quality, but they cause the change of the AQI value with the meteorological effects on PM_{10}, SO_2, CO and O_3, which harm human health.

In this study, values of AQI were predicted by means of ANN and k Nearest Neighbors (kNN), considering parameters of PM_{10}, SO_2, CO, O_3, temperature, humidity, pressure and wind speed. PM_{10}, SO_2, CO and O_3 values were obtained from Istanbul-Besiktas Air Quality Measuring Station of the Turkish Ministry of Environment and Urbanization. Temperature, humidity, wind speed and pressure data of the Besiktas district were taken from www.timeanddate.com First of all, the “AQI Calculator” (AQI Calculator, 2018) application was used to obtain the corresponding values of the AQI. The accuracy of the results obtained by the “AQI Calculator” application has been tested and verified with the current date parameters. The values obtained from this were confirmed by the available data on the National Air Quality Monitoring Network of the Ministry of Environment and Urbanization. AQI is unitless. Then, the AQI values obtained were classified as 1 to 5 mathematically. At the beginning 124 data for 2015, 2016, 2017 and 2018 January were used as data training data. The first 14 days of raw data for January 2019 were used as the test data. The training function and the number of neurons, which give the correct results at the highest rate and in the shortest time, have been applied to the data to be tested. The data to be tested were the first 14 days of January, March and December of 2018 and January of 2019. It was aimed by ANN and kNN to correctly predict the class of these three different 14 day data. Finally, the predicted and the mathematical classification results were compared.

2. Scientific Study Review

In a study conducted by Barrero et al. (2015) the PM_{10} concentrations of 2005-2012 were classified in 43 stations in Basque Region Air Quality Monitoring Networks. In this classification study, there were 5 groups with different PM_{10} levels. In warm and cold periods, the average daily PM_{10} values were used on weekdays and weekends. The autocorrelation functions of the daily and hourly PM_{10} time series were used to characterize changes. A methodology based on k-cluster was proposed. In the classification herein, it was found that, for PM_{10} concentration levels, monitoring stations that distinguish only the rural ones provide a better separation compared to the conventional categorization. The Evol-Ps based classification was found to be the most effective in comparing PM_{10} levels among groups. The results showed that the time series of pollutant concentrations contained sufficient information to provide an objective classification of the monitoring stations in an Air Quality Monitoring Network.

A mathematical framework was presented to formulate a Cumulative Index based on the individual concentration of the four major contaminants (SO_2, NO_2, $PM_{2.5}$ and PM_{10}) in a study by Saxena et al. (2017). Then, a classifier based on supervised learning algorithm was proposed. This classifier uses the Support Vector Machine to classify two types of air quality. The efficacy of the classifier was tested on the actual data in the Calcutta, Delhi and Bhopal regions. As a result, the classifier had been shown to perform well in determining air quality. Cumulative Index values and classification results obtained from the controlled learning model were found to be at the same level.

In a study by Karatzasetal. (2017) computational intelligence method was used to model the air pollution to the Gdansk Region in Poland. The prediction problem was analyzed by both classification and regression algorithms. Artifical NN based model that allows the use of the ensemble method was presented for the whole region. According to the results, the correlation coefficient between predicted and the hourly PM_{10} concentration measurements was 0.81 and the agreement index was 54%, indicating that the method used had a good performance. It was also concluded that this model could support the provision of air quality predicts on a company basis.

Alkasassbeh et al. (2013) developed a non-linear and Autoregressive ANN model (ANNARX) to predict both $PM_{10}$and Total Hanging Particles (TSP) in Jordan. The BP learning algorithm was used to train this model. The data set collected around a cement factory between 2006 and 2007 was used by 8 monitoring stations. In the proposed ANN model, temperature, relative humidity and wind speed data were used as input data. According to the experimental results, ANNARX can provide good modeling results using a limited number of measurement data.
In a study conducted by Nejadkoorki et al. (2012) an ANN model was proposed in Tehran to predict the 24-hour PM_{10} concentrations. It was aimed to predict the next day by using data from previous days of PM_{10} concentrations. Data from different stations were used between 2001 and 2009. Data of meteorological and gas pollutants from different air quality monitoring stations and meteorological areas were used as input data. Feed-forwardback propagation neural network, hyperbolic tangent sigmoid activation function and the Levenberg-Marquardt optimization methods were applied together. They reported that the PM_{10} prediction was found to be successful at 0.83 percent in all regions. In addition, it was concluded that Neural Networks developed for each monitoring region in the air quality monitoring network could compensate some of the missing data in neighboring monitoring zones.

In their study, Rahman etal. (2015) presented a time series model using data from three different stations between 2000 and 2009 in order to predict the Air Pollution Index. The Box–Jenkins approach of seasonal auto regressive integrated moving average (ARIMA), ANN and three models of fuzzy time series (FTS) had been compared by using the mean absolute percentage error, mean absolute error, mean square error and root mean square error. As a result, it was reported that the ANN showed the best result in predicting air pollution.

In the past, the most important pollutant sources for Besiktas were vehicle traffic and fossil fuels. SO_2 level was high due to fossil fuels. Nowadays, SO_2 level has decreased with the use of natural gas in heating. However, the rapidly increasing number of vehicles continues to cause a high amount of PM.

In this study, mathematical and predicted AQI classifications compared. Temperature, humidity, pressure, wind speed, PM_{10}, SO_2, CO and O_3 parameters were taken into account for prediction of AQI values. For comparison, 2 types of studies were conducted for 3 different periods. The study periods were January 2019, March 2018 and December 2018 and the first 14 days of these periods were recorded. The first 14 days of January 2019 was normalized and the study was performed to see if the raw data and the normalized data would affect the prediction result. Thus, the success of ANN was measured. It is thought that this study on AQI classification will contribute to the literature. PM_{10}, SO_2, CO and O_3 values were obtained from Istanbul-Besiktas Air Quality Measuring Station of the Turkish Ministry of Environment and Urbanization. Temperature, humidity, wind speed and pressure data of the Besiktas district were taken from www.timeanddate.com.

3. Recommended Method

3.1. Artificial Neural Networks and Algorithm

In this section, firstly ANN was introduced that was developed in 1943 by McCulloch and Pitts. Then, a study was conducted on how to obtain the predicted classification of the AQI value corresponding to the entered data corresponding to PM_{10}, SO_2, CO, O_3, temperature, humidity, pressure and wind speed. The ANN is a logical software developed to perform the basic functions of the human brain, such as learning, remembering and generalizing of the brain by imitating the working mechanism of the brain. They are parallel operation systems that are connected to each other in a network and are called a neuron and are composed of a series of processing units. It is a calculation technique inspired by the nervous system of the human brain. It is a type of processing which can be used in non-linear and wide application area. They are generally trained for classification purposes or using some learning algorithms to make appropriate decisions using measurement data collected in the application. Neural networks have two main features that make it a very useful tool for predicting and solving modeling problems. These are learning and generalization skills (Alkasassbehet al. 2013, Artificial Neural Network, 2019).

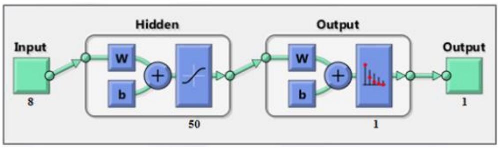

Figure 1. Neural Network (Neural Network Figure, 2019)

The neural network in Figure 1 contains 8 input data, 50 hidden neurons and 1 output data. The model needs to learn to produce the correct output. We do this by training weights. The “W” in the diagram is the abbreviation of the weights, and “b” is the abbreviation of the bias unit. The most common types of neural networks used in classification are feed-forward and feed-back networks. The feed-forward algorithm is an algorithm used to calculate the output vector from the input vector. Only a feed-forward algorithm is used when using a trained neural network (Feed-forward Back Propagation Neural Network, 2019). Feed-forward ANN have a one-way flow of information. In this network model, information from the input layer is transmitted to the hidden layer. Output value is determined by processing information from hidden and output layers.

In this study, computer simulations were carried out on MATLAB environment. In the study, feed-forward back-propagation neural network was used as ANN algorithm. The result values were obtained by 3-fold cross validation.

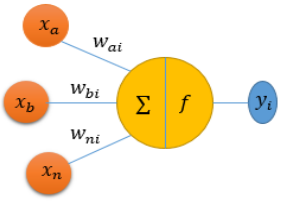

Figure 2. Feed-forward ANN (Feed-forward Artificial Neural Network Figure, 2019)

Summation function for feed-forward ANN as in equation 1.

$$I_i = \sum_b W_{bi} x_b \qquad (1)$$

The transfer function is as in equation 2.

$$y_i = f(I_i) \qquad (2)$$

Figure 2 shows a feed-forward ANN. Parameters such as thresholds and weights in conventional feed-forward neural networks must be updated through gradient-based learning algorithms (Baran, 2017). In order to obtain a special functional characteristic in response to an input set, the ANN algorithm based on the principle of adjusting the weights to generate outputs is called the back propagation algorithm (Back propagation Algorithm, 2018).

3.2. k Nearest Neighbors and Algorithm

k Nearest Neighbors (kNN) algorithm is one of the most used machine learning algorithm. It can uses for classification. Machine Learning makes inferences from ready data by using statistical and mathematical methods. kNN is a non-parametric learning algorithm. k nearest neighbors is an algorithm that stores all existing cases and classifies new cases by distance function criteria. Instead of learning the training data, it memorizes the data in training. A k value is determined in the operation of the algorithm. k is the number of elements to look at. When a value is reached, the nearest k element is taken and the distance between the value is calculated. If k is 1, then the case is simply assigned to its nearest neighbor. Historically, the optimal k values for most data sets were between 3-10. The Euclidean function is usually used for distance calculation. As an alternative to Euclidean function, Manhattan, Minkowski and Hamming functions can be used. Equations of these distance functions are shown in equations 3-6 (Baran, 2019; kNN1, 2019; kNN2, 2019).

$$\text{Euclidean} = \sqrt{\sum_{i=1}^{k}(x_i - y_i)^2} \quad (3)$$

$$\text{Manhattan} = \sum_{i=1}^{k} |\, x_i - y_i| \quad (4)$$

$$\text{Minkowski} = \left(\sum_{i=1}^{k}(|\, x_i - y_i|)^q\right)^{1/q} \quad (5)$$

$$\text{Hamming} = D_H = \sum_{i=1}^{k} |\, x_i - y_i| \quad (6)$$

If x = y, D = 0, If x ≠ y, D = 1,

3.3. Training Data Set for ANN and kNN

PM_{10}, SO_2, CO and O_3 parameters were used to construct the training data set. These valueswere obtained from Istanbul-Besiktas Air Quality Measuring Station of the Turkish Ministry of Environment and Urbanization. The values of these parameters are daily average values. The AQI values were obtained by using the “AQI Calculator” application. The obtained AQI values were confirmed by using the current data obtained from the National Weather Monitoring Network website of the Turkish Ministry of Environment and Urbanization. The above mentioned parameters are monitored hourly (Directorate of Environment and Urbanization National Air Quality Monitoring Network, 2018; Koçak, 2018). Afterwards, temperature, humidity, wind speed and pressure data of the Besiktas district from January 2015, 2017, 2018 were taken from the site www.timeanddate.com (Besiktas, 2015, 2016, 2017, 2018, 2019 temperature, pressure, wind speed, humidity values, 2019). For these parameters data of 12:00 o'clock were used. Thus, 8-column training data consisting of PM_{10}, SO_2, CO, O_3, temperature, humidity, pressure and wind speed were obtained as input data. Similarly, AQI values consisting of 1 column were used as output data. The training data consisted of 124 lines and 124 of the AQI data were classified mathematically between 1 and 5. As a result of this process, 1st class consisted of 0, 2nd class consisted of 59, 3rd class consisted of 36, 4th class consisted of 22, 5th class consisted of 7 data.

Using the raw data, the test data of 14 lines (first 14 days) of 3 different periods were studied. Training data consist of data for 2015, 2016, 2017, 2018 years and January months. It consists of the total of 31-day data. The test data belong to the first 14 days of January, March and December of 2018 and January of 2019. PM_{10}, SO_2, CO and O_3 values are taken from the National Weather Monitoring Network web site of the Ministry of Environment and Urbanization. Temperature, humidity, pressure and wind speed data are taken from the website www.timeanddate.com.

3.4. AQI Calculation and Limit Values

AQI is an index of daily air quality. Indicates how clean or dirty the air is. The purpose of measuring the AQI is to take precautions if the air quality value is at levels that threaten human health. Many countries have their own AQI standards and scales. In this study, “AQI Calculator” application was used to calculate the AQI value. The values of the parameters PM_{10}, SO_2, CO and O_3were entered into this application and AQI values were calculated. Calculated AQI values were compared with the current data obtained from the National Air Monitoring Network of the Ministry of Environment and Urbanization, and it was confirmed that “AQI Calculator” had the right calculation. In this study, the determined AQI limit values and the assigned class values were presented in Table 1.

Table 1. AQI limit values and assigned class

0-20	21-40	41-60	61-80	81-....
1	2	3	4	5

4. Case Study

4.1. ANN Study with Raw Data

Twelve different training functions were used for the classification with ANN. These were trainlm, trainbr, trainbfg, trainrp, trainscg, traincgb, traincgf, traincgp, trainoss, traingdx, traingdm and traingd functions. Test times and test accuracy rates of these training functions were compared. These values are obtained as in Table 2. The number of hidden neurons used in this initial comparison step was chosen to be 50.

Table 2. Test Times and Test Accuracy values at different training functions (raw data-50 hidden neuron)

Training function	Test Time (s)	Test Accuracy (%)	Training function	Test Time (s)	Test Accuracy (%)
trainlm	0.5017	71.43	traincgf	0.4678	78.57
trainbr	20.642	78.57	traincgp	0.2577	64.29
trainbfg	1.7101	50.00	trainoss	0.3733	57.14
trainrp	0.2516	57.14	traingdx	0.4758	78.57
trainscg	0.2914	71.43	traingdm	0.1319	0.00
traincgb	0.4238	78.57	traingd	0.2835	57.14

According to these results, when the test accuracy ratios with respect to ANN training functions were examined, it was seen that the highest value was realized in trainbr, traincgb, traincgf and traingdx training functions with an accuracy rate of 78.57%. However, since the trainbr training function can reach this solution in a very long time period of 20.642 seconds, it had not be taken into consideration in the prediction studies with test data. Therefore, studies were carried out in different secret neuron numbers for traincgb, traincgf and traingdx training functions. As a result of the studies, the values presented in Table 3 were obtained.

As can be seen from Table 3, in the prediction study conducted for the first 14 days data of January 2019, the highest test accuracy rate was realized as 0.2008 second in 20 hidden neuron counts with traincgf training function.

Table 3. Test Times and Test Accuracy values at different training functions and different neuron numbers (raw data-January 2019)

Training functions/ hidden neurons	Test Time (s)	Average Test Accuracy (%)	Training functions/ hidden neurons	Test Time (s)	Average Test Accuracy (%)
traincgb -20	0.16925	66.67	traincgb -80	0.25113	59.52
traincgf -20	0.22008	80.95	traincgf -80	0.33528	54.76
traingdx-20	0.24645	59.52	traingdx-80	0.27833	38.10
traincgb -40	0.22223	66.67	traincgb -100	0.49128	59.52
traincgf -40	0.25198	73.81	traincgf -100	0.40425	52.38
traingdx-40	0.12865	45.24	traingdx-100	0.25785	40.47
traincgb -60	0.27855	76.19	traincgb -124	0.42030	50.00
traincgf -60	0.24698	64.29	traincgf -124	0.41088	42.86
traingdx-60	0.29123	40.48	traingdx-124	0.48800	38.29

Since this data is the data obtained by 3-fold cross validation. The prediction of 80.95 corresponds to a prediction of 11 to 12 data on average. For this reason, the program was run again. The result was 85.71% corresponding to 12 correct predictions. According to the results of traincgf-20 training function and neuron number, ANN classification was performed between January 14, 2019 and January 01, 2019. The predicted classification values by ANN, 8 input values, calculated AQI values, and the mathematical classification values corresponding to the value of this AQI were as in Table 4.

There were two dates that cannot be accurately predicted. The first one was calculated as 1st class mathematically on January 02, 2019, while it was predicted as 2nd class by ANN. The AQI value calculated on this date was 18. Since the value of 18 was very close to the limit value of 20, it was thought that ANN could not predict this correctly. The second false guess made by ANN was realized on January 06, 2019. Calculated AQI value was 2, while ANN was calculated as 3rd class. The value of the AQI calculated on this date was 34.

A similar study was conducted by ANN between March 14, 2018 and March 01, 2018 on the prediction of AQI classification by ANN. Studies of traincgb, traincgf and traingdx training functions (giving the highest test accuracy results for 50 neurons count) were performed in different numbers of secret neurons.

As a result of this study, “Test Times” and ”Average Test Accuracy” values were as in Table 5. In the AQI classification prediction study conducted for the first 14 days of March 2018, the highest test accuracy rate was obtained with 71.43% and traingdx-40 with 0.6257 second test period. Accordingly, the results obtained with traingdx-40 training function and number of hidden neurons in the ANN algorithm for the prediction of AQI values of the first 14 days of March 2018 were as in Table 6.

Table 4. Parameter values, calculated mathematical classification and ANN predicted classification values in the period of January 14, 2019- January 01, 2019

PM$_{10}$ (µg/m^3)	SO$_2$ (µg/m^3)	CO (µg/m^3)	O$_3$ (µg/m^3)	Temperature (^{0}C)	Humidity (%)	Pressure (mbar)	Wind Speed (km/h)	AQI values	M. AQI Class	ANN P. Class
31,2	8.7	596.5	3.5	9	74	1002	21	31	2	2
19,6	5.5	405.8	24.46	6	68	1010	15	24	2	2
17,0	3.2	490.7	23.93	6	86	1013	24	25	2	2
19,3	4.4	595.1	20.21	9	93	1015	22	30	2	2
26,4	10	727.3	19.53	12	67	1009	25	37	2	2
39,9	7.6	541.1	18.1	7	55	1014	34	40	2	2
25,0	3.1	605.7	15.38	3	60	1019	27	30	2	2
25,1	4.4	425.9	11.61	3	66	1019	26	25	2	2
34,1	5.3	469.6	15.5	4	70	1010	14	34	2	3
9,20	3.2	309.0	26.27	5	84	1013	20	26	2	2
14,6	3.9	370.7	20.94	4	72	1023	23	21	2	2
15,5	4.1	293.1	21.28	9	72	1011	18	21	2	2
18,0	3.9	361.2	16.13	8	79	1011	16	18	1	2
19,1	6.5	269.0	23.78	10	60	1020	25	24	2	2

AQI= AirQuality Index, W.S.=Windspeed, M.=Mathematically P.= Prediction

Table 5. Test Times and Test Accuracy values at different training functions and different neuron numbers (raw data-March 2018)

Training functions/ hidden neurons	Test Time (s)	Average Test Accuracy (%)	Training functions/ hidden neurons	Test Time (s)	Average Test Accuracy (%)
traincgb -20	0.1358	40.47	traincgb -80	0.4652	23.81
traincgf -20	0.2450	50.00	traincgf -80	0.6876	35.71
traingdx-20	0.2163	66.66	traingdx-80	0.2692	69.05
traincgb -40	0.2728	16.67	traincgb -100	0.2108	42.86
traincgf -40	0.2681	50.00	traincgf -100	0.2308	28.57
traingdx-40	0.6257	71.43	traingdx-100	0.5267	66.67
traincgb -60	0.2825	24.48	traincgb -124	0.2088	42.86
traincgf -60	0.3653	42.86	traincgf -124	0.5960	35.71
traingdx-60	0.3215	64.29	traingdx-124	0.5198	42.85

Table 6 shows the temperature, humidity, pressure, wind speed, PM$_{10}$, SO$_2$, CO and O$_3$ values measured between March 14, 2018 and March 01, 2018, the values of the AQI calculated according to these values, the mathematical classification values and the classification made by the ANN algorithm.

For the first 14 days of March 2018, 71.43% performance was achieved. This percentage means that only 10 of the 14 lines were predicted correctly. When Table 6 was examined, the calculated AQI value (51) on March 01, 2018 was placed mathematically in the third class and it was placed in the 1st class in the prediction made by ANN. Three incorrect predictions made by ANN were made on March 08, 2018, March 10, 2018 and March 12, 2018.

Table 6. Parameter values, calculated mathematical classification and ANN predicted classification values in the period of March 14, 2018- March 01, 2018

PM$_{10}$ (µg/m^3)	SO$_2$ (µg/m^3)	CO (µg/m^3)	O$_3$ (µg/m^3)	Temperature (^{0}C)	Humidity (%)	Pressure (mbar)	Wind Speed (km/h)	AQI values	M. AQI Class	ANN P. Class
58.3	4.3	528.3	89.6	18	39	1011	27	90	5	5
69.0	9.2	759.1	83.59	20	38	1013	17	84	5	5
63.5	21	1066	44.88	18	36	1015	9	63	4	5
46.6	8.5	769.0	50.08	17	34	1014	8	50	3	3
56.6	7.7	798.2	52.73	15	38	1015	14	57	3	5
38.4	4.0	539.9	65.53	14	51	1009	27	66	4	4
82.2	4.7	645.4	105.4	13	84	1007	11	10	5	2
80.6	5.1	558.9	59.5	12	74	1017	5	81	5	5
35.6	2.5	700.3	58.24	11	83	1013	9	58	3	3
25.2	0.8	455.7	57.22	17	41	1011	9	57	3	3
43.3	1.9	569.7	46.06	17	50	1008	27	46	3	3
49.1	2.8	677.6	43.02	16	55	1014	20	49	3	3
55.4	22	779.5	43.7	9	48	1019	7	55	3	3
32.6	2.7	472.1	50.56	7	79	1015	17	51	3	1

AQI= AirQuality Index, W.S.=Windspeed, M.=Mathematically P.= Prediction

The results of the study conducted with traincgb, traincgf and traingdx training functions between December 14, 2018- December 01, 2018 and in different secret neuron numbers are as in Table 7. "Test Times" and "Average Test Accuracy" values are shown in the table.

Table 7. Test Times and Test Accuracy values at different training functions and different neuron numbers (raw data-December 2018)

Training functions/ hidden neurons	Test Time (s)	Average Test Accuracy (%)	Training functions/ hidden neurons	Test Time (s)	Average Test Accuracy (%)
traincgb -20	0.1557	71.43	traincgb -80	0.2876	40.47
traincgf -20	0.1799	59.52	traincgf -80	0.2480	56.87
traingdx-20	0.2171	61.91	traingdx-80	0.2785	57.15
traincgb -40	0.17875	78.57	traincgb -100	0.2936	52.38
traincgf -40	0.2157	51.06	traincgf -100	0.3446	42.86
traingdx-40	0.1735	42.86	traingdx-100	0.4263	64.29
traincgb -60	0.2374	38.09	traincgb -124	0.3922	50.00
traincgf -60	0.2677	71.43	traincgf -124	0.4730	50.00
traingdx-60	0.1223	71.43	traingdx-124	0.1919	14.29

The highest test accuracy obtained by using the test data of these dates is in traincgb training function and in 40 hidden neuron count with 0.17875 second test times and 78.57% rate. The ratio of 78.57% means that 11 of the classification data of the test data are correct. Accordingly, the comparison of ANN classification values and mathematical classification values for traincgb-40 was as in Table 8.

Table 8. Parameter values, calculated mathematical classification and ANN predicted classification values in the period of December 14, 2018 - December 01, 2018

PM_{10} (µg/m^3)	SO_2 (µg/m^3)	CO (µg/m^3)	O_3 (µg/m^3)	Temperature (ºC)	Humidity (%)	Pressure (mbar)	Wind Speed (km/h)	AQI values	M. AQI Class	ANN P. Class
40.5	9.7	452.1	4.35	4	82	1029	7	40	2	3
32.7	7.97	462.4	11.03	7	55	1023	16	33	2	2
29.9	3.97	563.6	10.5	10	75	1010	19	30	2	2
26.3	1.85	479.0	15.27	14	57	1013	22	26	2	2
29.2	4.72	492.8	9.43	10	79	1017	6	29	2	2
52.0	11.6	636.0	5.47	10	68	1020	6	52	3	3
38.1	9.04	519.2	7.44	8	74	1017	4	38	2	2
23.2	2.37	290.9	20.39	10	84	1012	6	23	2	2
19.9	2.01	373.8	16.16	11	79	1011	9	19	1	1
17.3	2.67	409.6	16.79	11	85	1015	5	17	1	2
25.2	2.72	338.1	19.2	8	92	1019	7	25	2	2
20.7	2.78	430.8	16.07	11	69	1022	6	21	2	2
18.6	3.17	350.9	16.0	11	77	1024	6	19	1	2
29.3	3.69	336.7	14.15	13	80	1018	11	30	2	2

AQI= AirQuality Index, W.S.=Windspeed, M.=Mathematically P.= Prediction

Table 8 shows the temperature, humidity, pressure, wind speed, PM_{10}, SO_2, CO and O_3 values measured for the first 14 days of December 2018, the values of the AQI calculated according to these values, the mathematical classification values and the classification made by the ANN algorithm. In the traincgb-40 training function and the number of neurons, 3 false predictions were made by ANN. These were on December 02, 2018, December 05, 2018 and December 14, 2018, and on these dates the AQI values are 19, 17 and 40, respectively. All three AQI values are close to the limit values. Because this proximity could not be predicted by ANN, the results were determined as a higher class.

4.2. ANN Study with Normalized Data

Training of ANN can be very slow when applied to raw data. Therefore, working with normalized data allows for faster results. For this purpose, all values between the lower and upper limit values of the eight parameters used in the study were normalized between 0 and 1 value. "Min-Max Normalization" technique was used to normalize raw data. The Min-Max technique normalizes data linearly. Minimum refers to the lowest value a data can receive while, maximum refers to the highest value that data can receive. It reduces the data to a value between 0 and 1. In this method, the largest and smallest values within a group of data are considered. All other data is normalized according to these values. The aim here is to normalize the smallest value to 0 and the maximum value to 1 and to spread all other data to the range 0-1. This equation is as in equation 7 (Yavuz et al. 2012).

$$x' = \frac{x_i - x_{min}}{x_{max} - x_{min}} \quad (7)$$

x'= normalized data,
x_i=input value
x_{max}= the largest number in the input set,
x_{min}= the smallest number in the input set,

Between the dates of January 14, 2019-January 01, 2019 traincgb, traincgf, traingdx training functions with different numbers of hidden neurons obtained as a result of the study, “Test Times” and “Average Test Accuracy” values as in Table 9. As can be seen from Table 9, in the prediction study for the normalized data of the first 14 days of January 2019 with ANN, the highest test accuracy rate was in traincgb-20, traincgb-40 and traincgf-60 training function and neuron counts. The test times for these functions were 0.2228, 0.2149 and 0.4227 second respectively. 85.71% performance was achieved. Table 10 was obtained by using the traincgb-40 function which predicts these times as soon as possible.

Table 9. Test Times and Test Accuracy values at different training functions and different neuron numbers (normalized data-January 2019)

Training functions/ hidden neurons	Test Time (s)	Average Test Accuracy (%)	Training functions/ hidden neurons	Test Time (s)	Average Test Accuracy (%)
traincgb -20	*0.2228*	*85.71*	traincgb -80	0.2061	50.00
traincgf -20	0.2009	64.29	traincgf -80	0.5704	71.43
traingdx-20	0.2146	78.57	traingdx-80	0.1307	7.14
traincgb -40	*0.2149*	*85.71*	traincgb -100	0.3794	64.29
traincgf -40	0.2114	35.71	traincgf -100	0.4474	42.86
traingdx-40	0.1630	35.71	traingdx-100	0.2904	50.00
traincgb -60	0.3894	78.57	traincgb -124	0.5455	35.71
traincgf -60	*0.4227*	*85.71*	traincgf -124	0.1926	21.43
traingdx-60	0.4299	50.00	traingdx-124	0.4734	35.71

When Table 10 is examined, it can be seen that there are two accurately unpredictable dates. The first one was calculated as 1st class on January 02, 2019, while it was predicted as 2nd class by ANN.

As in the study conducted with raw data, the calculated AQI value was 18 and it was thought to be due to the fact that ANN result cannot be predicted because it is close to the limit value of 20. The second false prediction made by ANN was realized on January 04, 2019. The calculated AQI value for this date was 21 and it was thought to be due to the fact that ANN cannot predict because it was close to the limit value of 20. In January 2019, 85.71% performance was achieved in the classification study conducted with both raw data and normalized data. This means that conducting the study with raw data will not adversely affect the outcome of ANN.

Table 10. Parameter values, calculated mathematical classification and ANN predicted classification values in the period of January 14, 2019- January 01, 2019 (normalized data)

PM_{10} (µg/m^3)	SO_2 (µg/m^3)	CO (µg/m^3)	O_3 (µg/m^3)	Temperature (ºC)	Humidity (%)	Pressure (mbar)	Wind Speed (km/h)	AQI values	M. AQI Class	ANN P. Class
0.25	0.38	0.31	0.00	0.62	0.56	0.20	0.39	31	2	2
0.13	0.22	0.21	0.51	0.48	0.45	0.38	0.23	24	2	2
0.11	0.11	0.25	0.50	0.48	0.78	0.44	0.46	25	2	2
0.13	0.17	0.30	0.41	0.62	0.91	0.49	0.41	30	2	2
0.20	0.44	0.37	0.39	0.76	0.44	0.36	0.49	37	2	2
0.34	0.33	0.28	0.36	0.52	0.22	0.47	0.72	40	2	2
0.19	0.10	0.31	0.29	0.33	0.31	0.58	0.54	30	2	2
0.19	0.17	0.22	0.20	0.33	0.42	0.58	0.51	25	2	2
0.28	0.21	0.24	0.29	0.38	0.49	0.38	0.21	34	2	2
0.03	0.11	0.16	0.56	0.43	0.75	0.44	0.36	26	2	2
0.08	0.14	0.19	0.43	0.38	0.53	0.67	0.44	21	*2*	*1*
0.09	0.15	0.15	0.43	0.62	0.53	0.40	0.31	21	2	2
0.12	0.14	0.18	0.31	0.57	0.65	0.40	0.26	18	*1*	*2*
0.13	0.27	0.14	0.50	0.67	0.31	0.60	0.49	24	2	2

AQI= AirQuality Index, W.S.=Windspeed, M.=Mathematically P.= Prediction

4.3. kNN Study with Raw Data

Studies were carried out in different neighborliness number. As a result of the studies, the values presented in Table 11 were obtained. As can be seen from Table 11, in the prediction study conducted for the first 14 days data of January 2019, the highest test accuracy rate was realized at 3, 5, 7 k values. The performance ratio in these k values was 92.86%. Of the fourteen days classification value, thirteen days were correctly estimated by the kNN algorithm.

Table 11. Accuracy Rate for different k Neighborliness Numbers (raw data-January 2019)

2019-January	**Classification by Neighborliness Number**			
M. AQI Class	***3***	***5***	***7***	***9***
2	2	2	2	2
2	2	2	2	2
2	2	2	2	2
2	2	2	2	2
2	3	2	2	2
2	2	2	2	2
2	2	2	2	2
2	2	2	2	2
2	2	2	2	2
2	2	2	2	2
2	2	2	2	2
2	2	2	2	2
1	2	2	2	2
2	2	2	2	2
Accuracy Rate (%)	85,71	92,86	92,86	92,86

A similar study was conducted by ANN between March 14, 2018 and March 01, 2018 on the prediction of AQI classification by kNN. For the first 14 days of March 2018, 28.57% performance was achieved. This percentage means that only 4 of the 14 days were predicted correctly. It can be seen Table 12.

Table 12. Accuracy Rate for different k Neighborliness Numbers (raw data-March 2018)

2018-March	**Classification by Neighborliness Number**			
M. AQI Class	***3***	***5***	***7***	***9***
5	4	3	3	3
5	4	4	4	4
4	4	4	3	3
3	2	2	2	2
3	2	2	4	4
4	2	2	2	2
5	4	4	4	4
5	4	4	4	3
3	2	2	2	2
3	3	3	2	2
3	3	3	3	3
3	2	2	2	2
3	2	2	4	4
3	3	2	2	2
Accuracy Rate (%)	28,57	21,43	7,14	7,14

Also, a study conducted for December 14, 2018- December 01, 2018 with different k values. The results of this study were as in Table 13. The highest accuracy rate was obtained at 3, 5, and 7 k values with 71,43%. The ratio of 71,43% means that 10 of the classification data of the test data are correct.

Table 13. Accuracy Rate for different k Neighborliness Numbers (raw data-December 2018)

2018-December	**Classification by Neighborliness Number**			
M. AQI Class	***3***	***5***	***7***	***9***
2	3	2	2	2
2	2	2	2	2
2	3	2	2	2
2	2	2	2	2
2	2	2	2	2
3	2	2	2	2
2	2	2	2	2
2	2	2	2	2
1	2	2	2	2
1	2	2	2	2
2	2	2	2	2
2	2	2	2	2
1	2	2	2	2
2	2	2	2	2
Accuracy Rate (%)	57,14	71,43	71,43	71,43

4.4. kNN Study with Normalized Data

As can be seen from Table 14, in the prediction study for the normalized data of the first 14 days of January 2019 with kNN, the highest test accuracy rate was 92,86%. There is only one accurately unpredictable dates.

Table 14. Accuracy Rate for different k Neighborliness Numbers (normalized data-January 2019)

2019-January	**Classification by Neighborliness Number**			
M. AQI Class	***3***	***5***	***7***	***9***
2	2	2	2	2
2	2	2	2	2
2	2	2	2	2
2	2	2	2	2
2	2	2	2	2
2	2	2	2	2
2	2	2	2	2
2	2	2	2	2
2	2	2	2	2
2	2	2	2	2
2	2	2	2	2
2	2	2	2	2
1	2	2	2	2
2	2	2	2	2
Accuracy Rate (%)	**92,86**	**92,86**	**92,86**	**92,86**

5. Results

In this study, it was aimed to predict the AQI values by ANN algorithm. For this purpose, eight parameters have been selected which may affect the AQI. These were PM_{10}, SO_2, CO, O_3, temperature, humidity, pressure and wind speed respectively. First of all, a data set was created for training purposes. Eight input data that consisting of 124 lines of PM_{10}, SO_2, CO, O_3, temperature, humidity, pressure and wind speed were created. On the other hand, training data consisting of AQI data were generated. The obtained 124 AQI data were classified mathematically between 1 and 5. The test data of 14 lines of the three different periods (the first 14 days of the test month) were studied with raw data. Training data was composed of 31-day data for 2015, 2016, 2017 and 2018 January. The data consisted of 124 lines. The test data were January, March and December 2018 and January 2019, respectively. The first 14-day data were used in the studies.The AQI values were obtained by “AQI Calculator” application. The results obtained were confirmed by using the current data obtained from the National Air Quality Monitoring Network website of the Ministry of Environment and Urbanization. The classification values obtained were classified mathematically between 1 and 5.

12 different training functions were used in the classification. In the first stage, 50 hidden neurons were studied to determine which training functions would find the most appropriate solution. As a result of the study, it was seen that the highest values were met by traincgb, traincgf and traingdx training functions with 78.77% test accuracy rate according to ANN training functions. However, in order to examine how these training functions would react in different numbers of neurons, the dependence of these 3 training functions on six different neurons was investigated. After the study, the predicted performance rate of ANN with raw data for January 2018 was 85.71%, for March 2018 was 71.43%, for December 2018 was 78.57% and with normalized data for January 2019 was 85.71%. The same performance rate was obtained for both the raw and normalized classification studies of the first 14 days of January 2019 data. This ratio was 85.71% performance which corresponds to 12 correct predictions.

The predicted performance rate of kNN with raw data for January 2019 was 92.86% with 5 k value, for March 2018 was 28.57% with 3 k value, for December 2018 was 71.43% with 5 k value. With normalized data for January 2019 was 92,86 % performance rate. When the accuracy rate obtained from ANN and kNN studies were compared, ANN with 92.86% in January 2018, kNN with 71.43% in March 2018 and kNN with 78.57% in December 2018 had higher accuracy rates. As a result of the studies carried out with both raw and normalized data for January 2019, ANN achieved 85.81% accuracy rate, whereas kNN yielded 92.86% performance with both raw data and normalized data.

Conflict of Interest

No conflict of interest was declared by the author.

References

Alkasassbeh, M., Sheta, A.F., Faris, H.,Turabieh, H., 2013. Prediction of PM10 and TSP Air Pollution Parameters Using Artificial Neural Network Autoregressive, External Input Models: A Case Study in Salt, Jordan. Middle-East Journal of Scientific Research. 14, 999-1009.

Artificial Neural Network, http://kod5.org/yapay-sinir-aglari-ysa-nedir/ [last access date: 02.02.2019].

AQI Calculator: [Online]. Available:https://app.cpcbccr.com/ccr_docs/AQI%20-Calculator.xls. [last access date: 01.12.2018].

Back propagation Algorithm: [Online]. Available:https://ab.org.tr/ab06/sunum/8.ppt. [last access date: 03.12.2018].

Baran, B., 2017. Yenilenebilir Enerji Kaynaklarını İçeren Mikro-şebeke Sistemlerin Akıllı Yönetimi, İnönü Üniversitesi, Fen Bilimleri Enstitüsü, Doktora Tezi, Malatya, Türkiye.

Baran B., 2019. Prediction of Air Quality Index by Extreme Learning Machines. International Conference on Artificial Intelligence and Data Processing (IDAP), Malatya, Turkey.

Barrero, M.A., Orza, J.A.G., Cabello, M., Cantón, L., 2015. Categorisation of air quality monitoring stations by evaluation of PM_{10} variability.Science of The Total Environment. 524–525: 225-236.

Besiktas, 2015, 2016, 2017, 2018, 2019 temperature, pressure, wind speed, humidity values. https://www.timeanddate.com/weather/turkey/besiktas/historic?month=1&year=2019. [Accessed 23 January 2019].

CO, O_3. [Online]. Available:https://www.epa.gov. [last access date: 12.01.2019].

Directorate of Environment and Urbanization National Air Quality Monitoring Network. [Online]. Available:www.havaizleme.gov.tr. [last access date: 27.11.2019].

Dursun, Ş., İbrahimova İ., 2014. Bakü Hava Kirlenmesinde SO_2'nin Rolü ve Meteorolojik Olaylarla İlişkisinin Araştırılması. Avrupa Bilim ve Teknoloji Dergisi. 1(3), 84-91.

Feed-forward Artificial Neural Network Figure. [Online]. Available:https://www.saedsayad.com/artificial_neural_network.htm.[last access date: 18.01.2019].

Feed-forward Back Propagation Neural Network. [Online]. Available:https://stackoverflow.com/questions/28403782/what-is-the-difference-between-back-propagation-and-feed-forward-neural-network.[last access date: 01.02.2019].

Kaplan, Y., Saray, U., Azkeskin, E., 2014. Hava Kirliliğine Neden olan PM10 ve SO2 maddesinin Yapay Sinir Ağı Kullanılarak Tahmininin Yapılması ve Hata Oranının Hesaplanması. Afyon Kocatepe Üniversitesi Fen ve Mühendislik Bilimleri Dergisi. 14(025201), 1-6.

Karacı, A., 2018. Development of PM2.5 Concentration Measurement Device for Intelligent City Air Tracking System And Asthma Diseases, Journal of Engineering Sciences and Design, 6(3), 418 – 425.

Karatzas, K., Katsifarakis, N., Orlowski, C., Sarzyński, A., 2017. Urban Air Quality Forecasting: A Regression and a Classification Approach.Asian Conference on Intelligent Information and Database Systems (ACIIDS 2017). 539-548.

kNN1 - k Nearest Neighbors Algorithm. [Online]. Available:https://medium.com/@ekrem.hatipoglu/machine-learning-classification-k-nn-k-en-yak%C4%B1n-kom%C5%9Fu-part-9-6f18cd6185d. [last access date: 15.11.2019].

kNN2 - k Nearest Neighbors Algorithm. [Online]. Available:https://www.saedsayad.com/k_nearest_neighbors.htm. [last access date:15.11.2019].

Koçak, E., 2018. Temporal Variation of PM10 and SO2 Concentrations of Aksaray Atmosphere: Conditional Bivariate Probability Function and Kmeans Clustering. Journal of Engineering Sciences and Design, 6(3), 471 – 478.

Kunt, F., Dursun, Ş., 2018. Konya Merkezinde Hava Kirliliğine Bazı Meteorolojik Faktörlerin Etkisi. Ulusal Çevre Bilimleri Araştırma Dergisi. 1(1), 54-61.

Menteşe, S., Tağıl, Ş., 2012. Bilecik'te İklim Elemanlarının Hava Kirliliği Üzerine Etkisi. Balikesir University The Journal of Social Sciences Institute. 15(28), 3-16.

Nejadkoorki, F., Baroutian, S., 2012. Forecasting Extreme PM10 Concentrations Using Artificial Neural Networks. Int. J. Environ. Res. 6(1), 277-284.

Rahman, N.H.A., Lee, M.H., Suhartono, L.M.T., 2015. Artificial neural networks and fuzzy time series forecasting: an application to air quality. Quality&Quantity International Journal of Methodology. 49(6), 2633–2647.

Neural Network Figure. [Online]. Available: https://blogs.mathworks.com/loren/2015/08/04/artificial-neural-networks-for-beginners/. [last access date:23.01.2019].

Saxena, A., Shekhawat, S., 2017. Ambient Air Quality Classification by Grey Wolf Optimizer Based Support Vector Machine. Journal of Environmental and Public Health. 12.

Sulfur Dioxide. [Online]. Available: www.environment.gov.au/protection/publications/factsheet-sulfur-dioxide-so2. [last access date:12.01.2019].

Tepe, A.M., Doğan, G., 2019. Investigation of Air Qualities of Four Cities Located on Southern Coast of Turkey. Journal of Engineering Sciences and Design, 7(3), 585 – 595.

Turgut, D., Temiz, İ., 2015. Time Series Analysis And Forecasting For Air Pollution in Ankara: A Box-Jenkıns Approach. alphanumericjournal. 3(2), 131-138.

XieY, Zhao B, Zhang L., Luo R., 2015. Spatiotemporal variations of $PM_{2.5}$ and PM_{10} concentrations between 31 Chinese cities and their relationships with SO_2, NO_2, CO and O_3.Particuology. 20, 141-149.

Yavuz, S., Deveci, M., 2012. İstatiksel Normalizasyon Tekniklerinin Yapay Sinir Ağın Performansına Etkisi. ErciyesÜniversitesi İktisadi ve İdari Bilimler Fakültesi Dergisi. 40(2), 167-187.